Numerical Optimisation: Background

Marta M. Betcke

m.betcke@ucl.ac.uk

Kiko Rullan

f.rullan@cs.ucl.ac.uk

Department of Computer Science,
Centre for Medical Image Computing,
Centre for Inverse Problems
University College London

Lecture 1

Mathematical opimisation problem

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{subject to} \quad \begin{array}{ll} c_i(x) = 0, & i \in \mathcal{E} \\ c_i(x) \geq 0, & i \in \mathcal{I} \end{array}$$

- $f : \mathbb{R}^n \to \mathbb{R}$: objective function
- $c_i : \mathbb{R}^n \to \mathbb{R}$: constraint function,
  $i \in \mathcal{E}$ equality constraints,
  $i \in \mathcal{I}$ inequality constraints.
- $x \in \mathbb{R}^n$: optimisation variable

Optimal solution $x^\star$ has the smallest value of f among all x which satisfy the constraints.

Geodesics are the shortest surface paths between two points.

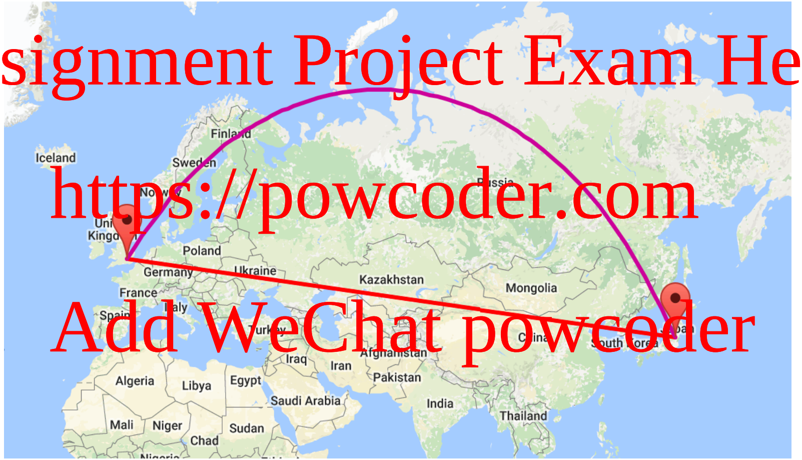


Figure: https://academo.org/demos/geodesics/

A very short and incomplete early history

Source `http://www.mitrikitti.fi/opthist.html`

- **Antiquity: geometrical optimisation problems**
  300 BC Euclid considers the minimal distance between a point a line, and proves that a square has the greatest area among the rectangles with given total length of edges
- **before Calculus of Variations: isolated optimization problems**
  1615 J. Kepler: optimal dimensions of wine barrel (with smallest variation of volume w.r.t. barrel parameters).[1]
  Early version of the secretary problem (optimal stopping problem) when he started to look for a new wife
  1636 P. Fermat shows that at the extreme point the derivative of a function vanishes. In 1657 Fermat shows that light travels between two points in minimal time.

[1] http://www.maa.org/press/periodicals/convergence/kepler-the-volume-of-a-wine-barrel-solving-the-problem-of-maxima-wine-barrel-design

A very short and incomplete early history cont.

Source http://www.mitrikitti.fi/opthist.html

- **Calculus of Variations**

  I. Newton (1660s) and G.W. von Leibniz (1670s) create mathematical analysis that forms the basis of calculus of variations (CoV). Some separate finite optimization problems are also considered

  1696 Johann and Jacob Bernoulli study Brachistochrone's problem, calculus of variations is born

  1740 L. Euler's publication begins the research on general theory of calculus of variations

A very short and incomplete early history cont.

Source http://www.mitrikitti.fi/opthist.html

- **Least squares**
  1806 A. M. Legendre presents the least square method, which also J.C.F. Gauss claims to have invented. Legendre made contributions in the field of CoV, too
- **Linear Programming**
  1826 J.B.J. Fourier formulates LP-problem for solving problems arising in mechanics and probability theory
  1939 L.V. Kantorovich presents LP-model and an algorithm for solving it. In 1975 Kantorovich and T.C. Koopmans get the Nobel memorial price in economics for their contributions on LP-problem
  1947 G. Dantzig, who works for US air-force, presents the Simplex method for solving LP-problems, von Neumann establishes the theory of duality for LP-problems

- 2 factories, F_i
- 12 retail outlets, R_j
- each factory F_i can produce up to a_i tones of a certain compound per week
- each retail outlet R_j has a weekly demand of b_j tones of the compound
- the cost of shipping of one tone of the compound from F_i to R_j is c_{ij}

Goal: what is the optimal amount to ship from each factory to each outlet which satisfies demand at minimal cost.

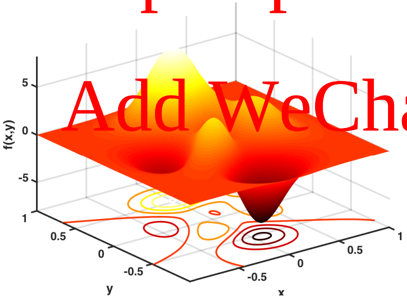
$$
\begin{aligned}
\min \quad & \sum_{ij} c_{ij} x_{ij} \\
\text{subject to} \quad \sum_{j=1}^{12} x_{ij} & \leq a_i, \; i = 1, 2 \\
\sum_{i=1}^{2} x_{ij} & \geq b_j, \; j = 1 \ldots 12 \\
x_{ij} & \geq 0, \; i = 1, 2, \; j = 1 \ldots 12.
\end{aligned}
$$

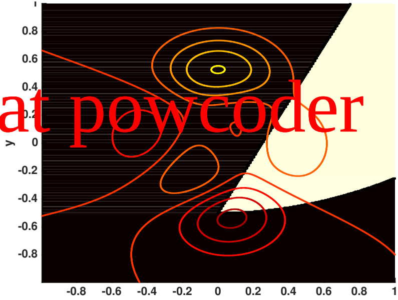
Linear programming problem because the objective function and all constraints are linear.

$$\min f(x,y)$$

$$\text{subject to} \quad -y + 2x - \frac{1}{2} \geq 0$$

$$y - \frac{1}{4}x^2 + \frac{1}{2} \geq 0.$$





A set $\mathbb{S} \subset \mathbb{R}^n$ is **convex** if for any two points $x, y \in \mathbb{S}$ the line segment connecting them lies entirely in $\mathbb{S}$

$$\alpha x + (1 - \alpha) y \in \mathbb{S}, \quad \forall \alpha \in [0, 1].$$

Examples:

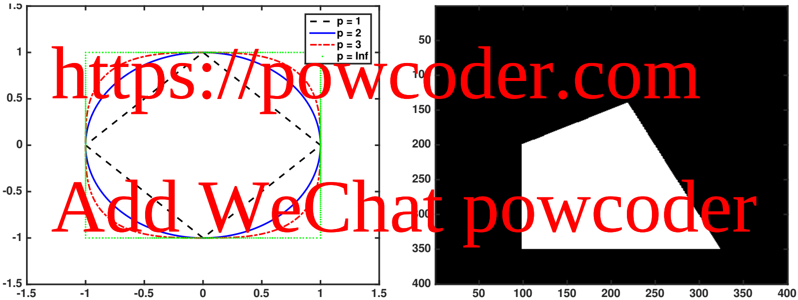


Figure: (a) unit ball $\{x \in \mathbb{R}^n : \|x\|_p \leq 1\},\ p \geq 1$; (b) polyheadron $\{x \in \mathbb{R}^n : Ax = b, Cx \leq d\}$

A function f is **convex** if

- its domain $\mathbb{S}$ is a convex set,
- for any two points $x, y \in \mathbb{S}$,

$$f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha) f(y), \quad \forall \alpha \in [0,1].$$

A function f is **strictly convex** if for $x \neq y$

$$f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha) f(y), \quad \forall \alpha \in (0,1).$$

A function f is **concave** if $-f$ is convex

Examples:

- linear function $f(x) = c^{\mathrm{T}} x + \alpha$, where $c \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$
- convex quadratic function $f(x) = x^{\mathrm{T}} H x$, where $H \in \mathbb{R}^{n \times n}$ symmetric positive (semi)definite

Classification of optimisation problems

- **convex** vs non-convex
- **smooth vs non-smooth**
- **constrained vs unconstrained**
- **linear vs quadratic vs nonlinear**
- **small vs large scale**
- **local** vs global
- stochastic vs deterministic
- discrete vs **continuous**

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a smooth function for which we can evaluate f and its derivatives at any given point $x \in \Omega \subseteq \mathbb{R}^n$.

Unconstraint optimisation problem

$$\min_{x \in \Omega \subseteq \mathbb{R}^n} f(x). \tag{1}$$

A point $x^\star$ is a **global minimiser** if

$$f(x^\star) \leq f(x), \ \forall x \in \Omega \subseteq \mathbb{R}^n.$$

A point $x^\star$ is a **local minimiser** if

$$\exists \mathcal{N}(x^\star) : f(x^\star) \leq f(x), \ \forall x \in \mathcal{N}(x^\star),$$

$\mathcal{N}(y)$ is a neighbourhood of y (an open set which contains y).

A point $x^\star$ is a **strict (or strong) local minimiser** if

$$\exists \mathcal{N}(x^\star) : f(x^\star) < f(x), \ \forall x \in \mathcal{N}(x^\star), \ x \neq x^\star.$$

Examples:

- $f(x) = 2$: every point is a (weak) local minimiser
- $f(x) = (x-2)^4$: $x^\star = 2$ is a strict local minimiser (also a global one)

- $f(x) = \cos(x)$: $x^\star = \pi + 2k\pi$, $k \in \mathbb{Z}$, are all strict local minimisers (but not strict global on $\mathbb{R}$)

A point $x^\star$ is an **isolated local minimiser** if

$$\exists \mathcal{N}(x^\star) : x^\star \text{ is the only local minimiser in } \mathcal{N}(x^\star).$$

Some strict local minimisers are not isolated e.g.

$$f(x) = x^4 \cos(1/x) + 2x^4, \quad f(0) = 0$$

has a strict local minimiser at $x^\star = 0$ but there are strict local minimisers (albeit $f(x_j) \geq x_j^4 > 0 = f(x^\star)$) at nearby points $x_j \to 0,\ j \to \infty$.

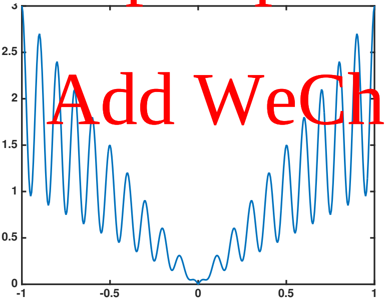
However, all isolated local minimiser are strict.

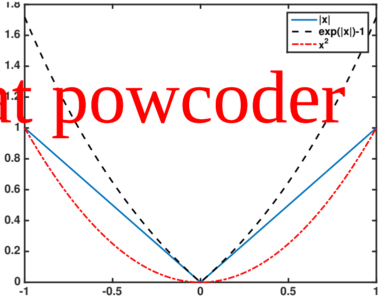
Difficulties with global minimisation:

$$f(x) = (\cos(20\pi x) + 2)|x|$$

has a unique global minimiser $x^\star = 0$, but the algorithms usually get trapped into one of the many local minima.

For convex functions, every local minimiser is also a global minimiser.





Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Then for $p \in \mathbb{R}^n$ we have

$$f(x+p) = f(x) + \nabla f(x+tp)^{\mathrm{T}} p$$

for some $t \in (0, 1)$.

If moreover f is twice continuously differentiable, we also have

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+tp) p dt.$$

and

$$f(x+p) = f(x) + \nabla f(x)^{\mathrm{T}} p + \frac{1}{2} p^{\mathrm{T}} \nabla^2 f(x+tp) p,$$

fore some $t \in (0, 1)$.

Theorem [1st order necessary condition]

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable in an open neighbourhood of a **local minimiser** $x^\star$, then $\nabla f(x^\star) = 0$.

Proof: [by contradiction]
Suppose that $\nabla f(x^\star) \neq 0$ and define $p = -\nabla f(x^\star)$. Note that $p^{\mathrm{T}} \nabla f(x^\star) = -\|\nabla f(x^\star)\|_2^2 < 0$. Furthermore, as ∇f is continuous near $x^\star$, there exists $T > 0$ such that

$$p^{\mathrm{T}} \nabla f(x^\star + tp) < 0, \quad t \in [0, T].$$

By Taylor's theorem, for any $\bar{t} \in (0, T]$ we have

$$f(x^\star + \bar{t}p) = f(x^\star) + \bar{t}p^{\mathrm{T}} \nabla f(x^\star + tp), \quad t \in (0, \bar{t}).$$

Hence $f(x^\star + \bar{t}p) < f(x^\star)$ for all $\bar{t} \in (0, T]$, and we have found a direction leading away from $x^\star$ along which f decreases which is in contradiction with $x^\star$ being a local minimiser.

We call $x^\star$ a **stationary point** if $\nabla f(x^\star) = 0$.

By Theorem [1st order necessary condition] any local minimiser is a stationary point. The converse is in general not true.

Theorem [2nd order necessary condition]

If $x^\star$ is a **local minimiser** of f and $\nabla^2 f$ exists and is continuous in an open neighbourhood of $x^\star$, then $\nabla f(x^\star) = 0$ and $\nabla^2 f(x^\star)$ is positive semidefinite.

Proof: [by contradiction]

By Theorem [1st order necessary condition] we have $\nabla f(x^\star) = 0$.
Assume $\nabla^2 f(x^\star)$ is not positive semidefinite. Then there exists a vector p such that $p^{\mathrm{T}} \nabla^2 f(x^\star) p < 0$, and because $\nabla^2 f$ is continuous near $x^\star$, there exists $T > 0$ such that $p^{\mathrm{T}} \nabla^2 f(x^\star + tp) p < 0$ for all $t \in [0, T]$.

By Taylor theorem, we have for any $\bar{t} \in (0, T]$ and some $t \in (0, \bar{t})$

$$f(x^\star + \bar{t}p) = f(x^\star) + \bar{t} p^{\mathrm{T}} \nabla f(x^\star) + \frac{1}{2} \bar{t}^2 p^{\mathrm{T}} f(x^\star + tp) p < f(x^\star).$$

We have found a decrease direction for f away from $x^\star$ which contradicts $x^\star$ being a local minimiser.

Theorem [2nd order sufficient condition]

Let $f : \mathbb{R}^n \to \mathbb{R}$ with $\nabla^2 f$ continuous in an open neighbourhood of $x^\star$. If $\nabla f(x^\star) = 0$ and $\nabla^2 f(x^\star)$ is positive definite, then $x^\star$ is a **strict local minimiser** of f.

Proof:

Because the Hessian $\nabla^2 f$ is continuous and positive definite at $x^\star$, we can choose a radius $r > 0$ so that $\nabla^2 f(x)$ remains positive definite for all x in an open ball $\mathcal{B}_2(x^\star, r) = \{y : \|y - x^\star\|_2 < r\}$.

For any nonzero vector $p \neq 0$, $\|p\|_2 < r$, $x^\star + p \in \mathcal{B}_2(x^\star, r)$ and

$$f(x^\star + p) = f(x^\star) + p^{\mathrm{T}} \nabla f(x^\star) + \frac{1}{2} p^{\mathrm{T}} \nabla^2 f(x^\star + tp) p \quad (2)$$

$$= f(x^\star) + \frac{1}{2} p^{\mathrm{T}} \nabla^2 f(x^\star + tp) p, \quad (3)$$

for some $t \in (0, 1)$.

Furthermore, $x^\star + tp \in B_2(x^\star, r)$ thus $p^{\mathrm{T}} \nabla^2 f(x^\star + tp) p > 0$ and therefore $f(x^\star + p) > f(x^\star)$.

2nd order sufficient condition guarantees a stronger statement than the necessary conditions (strict local minimiser).

A strict local minimiser may fail to satisfy the sufficient conditions:

$$f(x) = x^4, \quad f'(x) = 4x^3, \quad f''(x) = 12x^2$$

$x^\star = 0$ is a strict local minimiser while $f''(x^\star) = 0$ thus it satisfies the necessary but not the sufficient conditions.

If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, any local minimiser $x^\star$ is also a global minimiser of f. If, f is also differentiable, then any stationary point $x^\star$ is a global minimiser.

Proof:
Suppose $x^\star$ is a local but not global minimiser. Then $\exists z \in \mathbb{R}^n : f(z) < f(x^\star)$. For all x on a line segment joining $x^\star$ and z i.e.

$$\mathcal{L}(x^\star, z) = \{x : x = \lambda z + (1-\lambda)x^\star, \quad \lambda \in (0,1]\},$$

by convexity of f we have

$$f(x) \leq \lambda f(z) + (1-\lambda) f(x^\star) < f(x^\star).$$

For any neighbourhood $\mathcal{N}(x^\star) \cap \mathcal{L}(x^\star, z) \neq \emptyset$, hence $\exists x \in \mathcal{N}(x^\star) : f(x) < f(x^\star)$ and $x^\star$ is not a local mininiser.

Proof: cont.

For the second part, we suppose that $x^\star$ is not global minimiser.

For all z chosen as before by convexity of f it follows

$$\begin{aligned}
\nabla f(x^\star)^{\mathrm{T}}(z - x^\star) &= \left.\frac{d}{d\lambda} f(x^\star + \lambda(z - x^\star))\right|_{\lambda=0} \\
&= \lim_{\lambda \to 0} \frac{f(x^\star + \lambda(z - x^\star)) - f(x^\star)}{\lambda} \\
&\leq \lim_{\lambda \to 0} \frac{\lambda f(z) + (1-\lambda) f(x^\star) - f(x^\star)}{\lambda} \\
&= f(z) - f(x^\star) < 0.
\end{aligned}$$

Hence $\nabla f(x^\star) \neq 0$ and $x^\star$ is not a stationary point.